發表於 2008 台灣華語文教學年會暨研討會，花蓮慈濟大學，2008 年 11 月 1-2 日

華語文能力測驗垂直等化研究[1]

國家華語測驗推動工作委員會　藍珮君

摘要

本研究旨在針對國家華語測驗推動工作委員會研發之華語文能力測驗(Test Of Proficiency-Huayu，簡稱 TOP)初等、中等以及高等測驗進行垂直等化研究，以瞭解初、中、高三等測驗的試題難度分佈情形。

本研究以參加華語文能力測驗預試的華語學習者為對象，研究工具為華語文能力測驗初、中、高三等測驗。使用的等化設計為共同題不等組設計法(Common-Item Nonequivalent Groups Design)，在編擬預試試卷時，分別於初等及中等測驗、中等及高等測驗放置相同的試題。資料分析方面，採用「試題反應理論」(Item Response Theory)的等化方法，回收考生作答反應後，使用 ConQuest 軟體，先後以同時估計法(concurrent calibration)、平均數標準差法(mean and sigma method)估計並連結每道試題的難度參數(item difficulty parameter)。

分析結果主要針對 TOP 初、中、高三等測驗試題難度分佈情進行探討。研究結果符合預期，多數初、中、高三等測驗的試題難度區分良好，大致呈現初等測驗較為簡單，中等測驗次之，高等測驗較難的趨勢。惟中等測驗和高等測驗的平均難度差距較小，有待研發人員持續關切。

本研究結果除提供實徵數據說明華語文能力測驗初、中、高三等測驗的試題難度區分情形外，未來將持續藉由共同題不等組設計的等化方法，將預試試題的難度參數估計值轉換至此一共同量尺。待試題品質良好且完成對應的預試題目用於正式考試時，便可估計出考生能力參數，透過在不同等級 TOP 測驗獲得相近能力參數考生的答對總題數，便能瞭解初、中、高三等測驗之間的分數對應情形，可進一步建立三等測驗的分數對照表，提供給華語教師或華語學習者參考。

關鍵字：共同題、華語文能力測驗、測驗等化

[1] 本研究之完成，特別感謝華測會研發組組長林玲英、同仁洪小雯、簡秀文及張秀娟小姐之協助。

緒論

國家華語測驗推動工作委員會(簡稱華測會)研發之華語文能力測驗(Test Of Proficiency-Huayu，簡稱 TOP 測驗)，為一種外語/第二語言的能力測驗，測驗對象為母語非華語之人士。TOP 測驗共分為基礎、初等、中等和高等，一共四個等級。

TOP 測驗自 2003 年 12 月舉辦第一次正式考試以來，至今已超過四千人到考，考生國籍遍布六十餘國家。此外，承蒙教育部國際文教處駐外代表處文化組，以及海外各地華語教學機構的大力協助與支援，從 2006 年開始，華測會也陸續在海外各國舉辦正式考試與試測(中央社，民 97a；中央社，民 97b)。2008 年更擴大至 14 國 16 個地區進行施測，至目前為止，本年度海外地區參加 TOP 測驗的考生人數已達 2825 人。

為因應世界各地學習華語的熱潮，提供型態更為豐富多元的華語能力測驗，華測會研發人員除了積極研發寫作、口語測驗，與適合 15 歲以下華語學習者的兒童測驗之外，亦持續針對現有的四等聽讀測驗內容進行審視，並進行相關研究，以確保測驗的公正性及維護考生的權益。甫於 2007 年進行的 TOP 初中高三等測驗水平等化研究[2]，就是為了瞭解相同等級不同試卷之間的測驗難度是否相近，此研究也獲得令人滿意的結果，提出實徵數據說明了 TOP 初中高三等測驗難度具有穩定性。

一直以來，TOP 測驗研發人員面臨的難題之一便是如何提出實徵數據說明 TOP 各等測驗的難度差距。針對 TOP 測驗的各等考試，雖已發展出相對應的詞彙量和能力說明以輔助華語教師與研發人員進行命題及審題工作；然而「如何證明初等測驗比中等測驗簡單？」、「高等測驗比中等難多少？」等問題都是研發人員亟欲釐清的。因此有了實徵研究成果支持 TOP 初中高三等測驗難度的穩定性後，研究人員更進一步著手垂直等化研究，探究 TOP 各等測驗試題難度的分佈情形。由於基礎測驗的題型和題數與初中高三等測驗較為不同[3]，以共同題(common item)進行垂直等化較為困難，因此本研究先針對題數和題型相近的初中高測驗進行垂直等化研究，試圖解答上述問題。

文獻探討

1. 測驗等化的定義

等化(equating)是指使用統計方法，將一測驗的分數轉換至另一測驗分數量尺，以比較兩個測驗分數關係的過程，其目的是為了校準試題難度的差異。而且，等化不受試題施測內容和受試者能力分佈的影響。等化具有下列幾項特性(Kolen & Brennan，1995；引自曾玉琳、王暄博、郭伯臣、許天維，民 94)：

[2] 即將發表於 2008 年 11 月 21 -23 日在美國佛羅里達州舉行之 ACTFL 年會。

[3] TOP 四等測驗題數及題型詳見華測會官網介紹 http://www.sc-top.org.tw/chinese/LR/test1.php。

(1) 對稱性(symmetry)：等化之轉換必須是可逆的，即由 X 測驗等化至 Y 測驗的結果，與由 Y 測驗等化至 X 測驗之結果相同。
(2) 相同試題規格(same specifications property)：兩測驗欲進行等化，測驗內容需皆測量相同之能力特質。
(3) 相等性(equity)：受試者不論接受 X 測驗或 Y 測驗，所測得結果並無差異。
(4) 團體不變性(group invariance property)：等化過程中受試者不論來自何種團體樣本，所轉換出來的結果均相同。

2. 測驗等化的種類

Hambleton 與 Swaminathan(1985)指出測驗等化可分為水平等化(horizontal equating)與垂直等化(vertical equating)兩種，介紹如後(引自張鈺卿，民 96)：

(1) 水平等化

水平等化是指對兩個以上測量相同特質、相同能力且難度相近的測驗，將其原始分數轉換至同一量尺的過程。水平等化常被應用在許多大型測驗中，例如：托福、GRE，以及基本學力測驗等考試，就有多種複本測驗，可以在一年中實施多次的考試，然後藉由水平等化的過程，將不同複本測驗的成績轉換為同一量尺以進行比較。

(2) 垂直等化

垂直等化是指對兩個以上測量相同特質、相同能力但難度不一的測驗，將其原始分數轉換至同一量尺的過程。此類測驗的受試者的能力通常是屬於不同年齡或年級，如美國的加州成就測驗(California Achievement Tests，CAT)、愛奧華基本技能測驗(Iowa Test of Basic Skills)等，就是透過垂直等化的方式，將測驗與測驗之間的分數進行連結。

3. 測驗等化設計

測驗等化設計是指收集等化資料的方法。一般常見的等化設計包括單組設計(single group design)、平衡對抗隨機組設計(counterbalanced equivalent groups design)、等群組設計(equivalent group design)、平衡不完全區塊設計(balanced incomplete block design)、試題預先等化設計，以及共同題不等組設計(Common-item nonequivalent group design)等(王寶墉，民 84；Kolen & Brennan，1995)。

本研究採用的是共同題不等組設計，作法是在兩個不同能力分佈受試者母群體 P 和 Q 中，分別隨機抽取受試者樣本 P1 和 Q1。其中，P1 受試者接受 X 測驗，Q1 受試者接受 Y 測驗。P1、Q1 兩樣本受試者群另外需接受共同題(common items)A 測驗的施測。共同題不等組設計經常使用於一測驗只能被施測一次的測驗形式。通常共同題於兩個樣本群體中的施測順序是相同的，以避免施測順序的影響(order effects)，且共同題的內容和難度與 X、Y 測驗相似。共同題不等組設計如表 1 所示(Kolen & Brennan，1995；引自曾玉琳，民 94)。

表 1　共同題不等組等化設計

受試者群	X 測驗	Y 測驗	共同題 A 測驗
P1	O		O
Q1		O	O

由於共同題不等組設計無須讓考生在短時間內作答二份試卷，不會造成考生生理上的負擔與疲勞；可於進行預試時，事先規劃好共同題，安置於欲進行等化的測驗中，在實際實施上較為可行，故本研究採用共同題不等組等化設計。

4. 試題反應理論等化方法

測驗等化方法是指測驗等化資料收集完畢，進行測驗等化之時，連結測驗量尺之間的方法。常用的試題反應理論等化方法包括：(1)同時估計法(concurrent estimation)；(2)分離估計法(separate estimation)：包含平均數法(mean method)、平均數與標準差法(mean and sigma method)、特徵曲線法(characteristic curve method) (Kolen & Brennan，1995)。以下分別介紹本研究使用的「同時估計法」以及「平均數標準差法」二種等化方法。

(1)同時估計法

同時估計法是藉由測驗等化設計與 IRT 電腦軟體所提供之功能作連結，將所有測驗之測驗資料同時進行試題校準，經由校準後，就能將所有測驗之受試者能力值與試題參數放置在相同量尺上。其主要的原理為：將測驗等化設計測驗題本中之試題參數估計值同時對應於相同能力量尺上(許天維，民 95)。簡言之，同時估計法是利用各測驗具有的共同試題，將所有測驗試題串在一起，並同時估計試題參數。

由於同時估計法所得到的試題參數量尺只有一種，減少了測驗間連結時產生的誤差。且國內外許多文獻亦證實，採用同時估計法能獲得較佳的精準度(Hanson & Beguin, 2002；Kim & Cohen, 1998；陳煥文，民 93；引自許天維，民 95)，因此本研究採用同時估計法。

(2)平均數與標準差法

平均數與標準差法是利用 X 測驗與 Y 測驗二個測驗共同題難度參數的平均數和標準差，計算出 X 測驗和 Y 測驗試題參數量尺線性轉換的斜率 α 和截距 β，再將測驗分數利用線性轉換至 Y 測驗分數對應的分數。計算模式如下(Kolen & Brennan, 1995)：

$$\alpha = \frac{\sigma(b_Y)}{\sigma(b_X)}$$

$$\beta = \mu(b_Y) - \alpha\,\mu(b_X)$$

其中，b 指的是難度參數；

$\mu(b_x)$和 $\mu(b_y)$是 X 測驗和 Y 測驗共同題參數的平均數；

$\sigma(b_x)$和 $\sigma(b_y)$是 X 測驗和 Y 測驗共同題參數的標準差。

研究假設與目的

綜上所述，本研究旨在探討使用共同題不等組設計進行垂直等化研究之華語文能力測驗初中高三等測驗，試題難度的分佈情形。研究假設如下：

初、中、高三等測驗的試題難度區分良好，大致呈現出初等測驗較為簡單，中等測驗次之，高等測驗較難的趨勢。

研究方法

1. 研究設計

本研究為瞭解初等、中等以及高等測驗的試題難度分佈情形，採用共同題等化設計(common items design)，於 2008 年 6 月份進行預試組卷時，分別在初等及中等、中等及高等測驗之中，放入相同的試題以進行垂直等化。華語文能力測驗總題數為一百二十題，Livingston(2004)建議超過 100 題的測驗，共同題題數至少為 20 題；此外，共同題宜為原測驗之迷你版本(mini form)，各題型試題比例宜與原測驗相仿。根據上述二原則，本次初等與中等測驗、中等與高等測驗的預試共同題題數各為 23 題。詳細分測驗與題型題數分佈，及共同題題數分佈如表 2 所示。

表 2　TOP 初、中、高等測驗共同題分佈情形[4]

等級		聽力理解測驗			詞彙語法測驗		閱讀理解測驗		
		單句	對話	段落	詞彙	語法	單句	材料	短文
初	原始題數	20	20	10	20	20	10	20	—
	共同題數	3	4	3	3	4	2	4	—
中	原始題數	15	20	15	10	20	10	10	20
	共同題數	6	8	6	6	7	5	4	4
高	原始題數	15	20	15	20	10	10	—	30
	共同題數	3	4	3	3	3	3	—	4

2. 研究參與者

本研究參與者為 2008 年 4 月到 5 月份海外施測非華裔考生[5]以及參加 6 月份 TOP 測驗預試之考生，考生可根據華測會建議的學習時數或詞彙量[6]，選擇適合的等級報考，各等測驗到考人數如表 3 所示。初等測驗有 229 人到考；中等測驗有 209 名考生；參加高等測驗的人數較少，有 118 人。合計參與本次垂直等化研究的考生共有 556 人。

[4] 初等測驗沒有短文題型；高等測驗沒有材料題型。

[5] 非華裔考生在此指的是母語非華語，且平常不與家人或朋友使用中文交談之考生。

[6] 初、中、高等測驗適用對象，請參見華測會官網 http://www.sc-top.org.tw/。

表 3　TOP 初、中、高等測驗到考人數一覽表

測驗等級	施測地區	人數
初等	台灣	186
	巴拉圭	14
	俄羅斯	19
	波蘭	10
中等	台灣	196
	巴拉圭	5
	俄羅斯	7
	波蘭	1
高等	台灣	114
	馬來西亞	2
	俄羅斯	2
合計		556

3. 資料分析

本研究使用 ConQuest 軟體進行「同時估計法」分析，採用試題作答理論的單參數模式(one parameter model)，估計 TOP 初、中、高三等測驗每道試題的難度參數估計值(β 值)以及每位考生的能力參數估計值[7](θ 值)。

但因同時估計法疊代(iteration)過程估計值收斂情形不佳，諮詢台師大測驗領域教授後，建議改採分離估計法。先個別估計初中高三等測驗試題難度參數，然後以中等測驗共同題難度參數作為參照標準，使用平均數與標準差法連結初等與高等測驗共同題參數，再轉換初等以及高等測驗的試題難度參數至此一量尺。最後依照試題難度參數估計值的大小，依序排列出各等測驗試題難度的分佈情形。

研究結果與討論

在試題反應理論中，難度參數是指平均數為 0，標準差為 1 的數值，若難度參數為正值表示題目較難；若為負值則表示題目較為簡單。由於進行平均數與標準差法等化時，研究者發現初中等和中高等分別有 9 題以及 2 題共同題存在試題差異功能(differential item functioning，簡稱 DIF)，有研究指出等化過程中若共同題的 DIF 試題所佔比例越高，對於受試者能力估計值估計精準度誤差越大(蔡良庭、施懿珊，民 94)。為避免此 11 題有 DIF 之共同題影響後續等化結果，故予以刪除。

難度參數等化結果如表 4 所示，無論包含共同題與否，三等測驗平均難度均為高等測驗最難，中等測驗次之，初等測驗最簡單，此一結果符合預期。而由不含共同題的數值可知，初等與中等測驗平均難度相差 2.162；中等測驗和高等測

[7] 試題反應理論(item response theory)中，能力參數(θ 值)代表測驗測得考生的能力程度(ability level)，在此可視為古典測驗理論的總分。

驗平均難度差距較小，為 0.525。此外，最簡單的試題在初等測驗，難度參數為 -4.316；最難的題目落在高等測驗，難度參數為 3.666。

表 4　TOP 初、中、高等測驗難度參數描述統計

	測驗等級	題數	最小值	最大值	平均數	標準差
含共同題	初等	111	-4.316	2.523	-1.938	1.226
	中等	109	-3.483	2.767	0.014	0.976
	高等	118	-2.129	3.666	0.590	0.897
不含共同題	初等	97	-4.316	2.523	-2.064	1.204
	中等	74	-2.316	2.767	0.098	0.888
	高等	97	-2.129	3.666	0.623	0.930

TOP 初、中、高三等測驗試題難度分佈情形如圖 1 所示，試題排列順序係依照難度參數，由難至易排列，每一個圓點代表一題。橫軸表示試題等級，「高等」表示該題為高等測驗試題；「中高共」表示該題為中等和高等測驗共同題，以此類推。從圖 1 可知，初等、中等以及高等測驗的試題難度分佈大體上有所區隔，難度參數最低的試題落在初等測驗，最難的則是高等測驗的題目。然而，中等和高等測驗的試題難度差異不夠明顯。

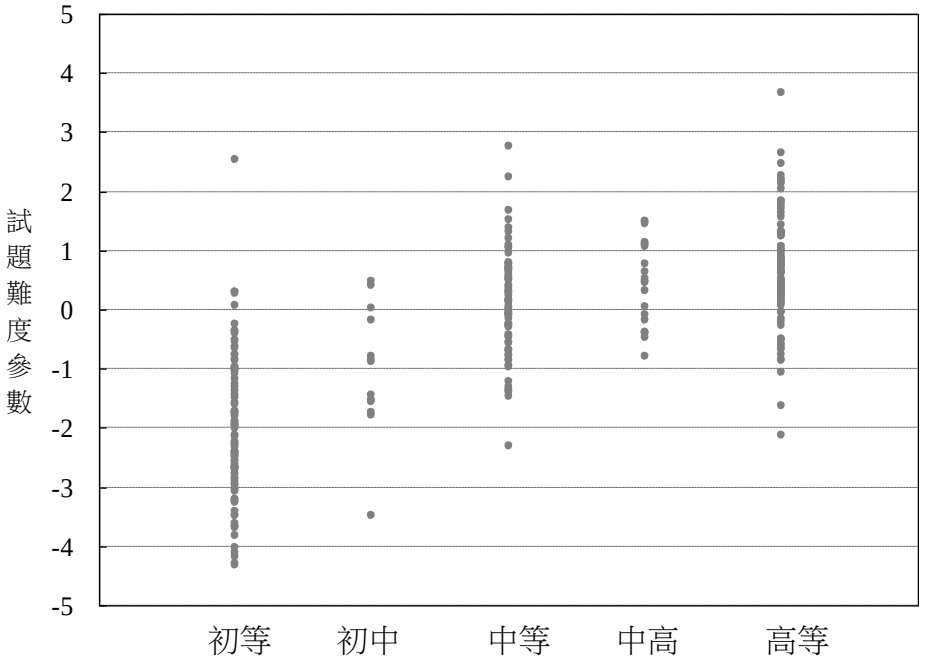

圖 1　TOP 初、中、高等測驗試題難度參數分佈圖

此一結果顯示，TOP 初、中、高三等測驗的試題分佈情形符合研究假設：高等測驗試題較難，中等測驗次之，初等測驗較為容易。惟中等測驗和高等測驗平均試題難度之差距稍嫌過小。若要明確區分初、中、高三個測驗等級，宜再提高高等測驗試題的難度，拉開與中等測驗的差距，如此三個等級測驗試題難度的分佈區塊就會更為明顯與清楚，更加符合測驗分等的目的。

至於考生能力參數的部分，由於部分共同題有 DIF，在進行試題難度參數等化時已經刪除，連帶造成不同等級考生原始題數不一(見表 4)，不適宜進行初中高三等測驗原始分數的對照。分數對照表的建立，仍有賴未來持續進行試題等化連結，將預試試題的難度參數估計值轉換至此一共同量尺。待試題品質良好且完成對應的預試題目用於正式考試時，便可估計出考生能力參數，以進一步瞭解初、中、高三等測驗之間的分數對應情形。

結論與建議

由垂直等化研究的結果，可以瞭解目前 TOP 初、中、高三等測驗試題分佈情形符合研究預期，沒有出現初等測驗試題平均難度比中等測驗難，或是高等測驗試題平均難度比中等測驗簡單的情形。可見華測會測驗研發人員對於 TOP 初中高三等測驗的試題難度分級與品質控管達到一定水準，考生可自行選擇適合的等級報考，以瞭解自身的華語文程度。然而，中等與高等測驗之間的難度區隔得不夠明顯，還需研發人員對此進行討論或調整，如：提高高等測驗的試題難度。

未來若要提高試題難度，除了參照詞彙量和語法結構之外，研究者認為測驗主題與情境的適度擴充是必要的。如果只侷限在生活、公眾或教育領域，沒有涉及到職場或更為專業的範疇，並深入該領域命題，即使使用較艱深的詞彙或語法，可能還是無法測得更為高層次的華語文能力。

本研究提出實徵數據說明 TOP 初中高測驗試題難度分佈情形，並首度建立起橫跨三等測驗的試題難度參數共同量尺，未來可將透過等化設計進行預試後的新題目難度連結至此一量尺。當這些試題運用在正式考試時，便能提供更為豐富的試題訊息予研發人員進行組卷，以維持各等測驗以及各測驗版本難度的穩定性，確保測驗的公正性並維護考生權益。

參考文獻

1. 中文部分

王寶墉(民 84)：**現代測驗理論**。台北市：心理出版社。

張鈺卿(民 96)。**BIB 與 NEAT 設計在不同年度測驗連結效果之比較**。國立台灣教育大學教育測驗統計研究所碩士論文，未出版。

推動繁體中文　台灣在泰舉辦第三屆華文測驗（民 97b 年 4 月 27 日）。**中央社**。民國 97 年 5 月 8 日，取自：http://tw.news.yahoo.com/article/url/d/a/080427/5/y38j.html

許天維(民 95)。**大型教育測驗等化設計及效果之研究**。國科會專題研究計畫。

曾玉琳(民 94)。**不同配置設計下測驗等化效果之模擬研究**。國立台中師範學院數學教育學系碩士論文，未出版。

曾玉琳、王暄博、郭伯臣、許天維(民 94)。**不同 BIB 設計對測驗等化的影響**。測驗統計年刊，第十三輯下期，209-229 頁。

蔡良庭、施懿珊(民 94)。**具差異試題功能之定錨試題對測驗等化之影響**。測驗統計年刊，第十三輯上期，95-109 頁。

駐紐約文化組將再舉辦三場華語文能力測驗（民 97a 年 4 月 11 日）。**中央社**。民 97 年 5 月 8 日，取自：http://tw.news.yahoo.com/article/url/d/a/080411/5/x3bj.html

2. 英文部分

Kolen, M.J. & Brennan, R.J. (1995). *Test Equating: Methods and Practices*. New York: Springer-Verlag.

Livingston, S. A. (2004). *Equating test scores*. Retrieved March 14, 2007, from http://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf